

# Python Machine Learning Labs

Good morning!

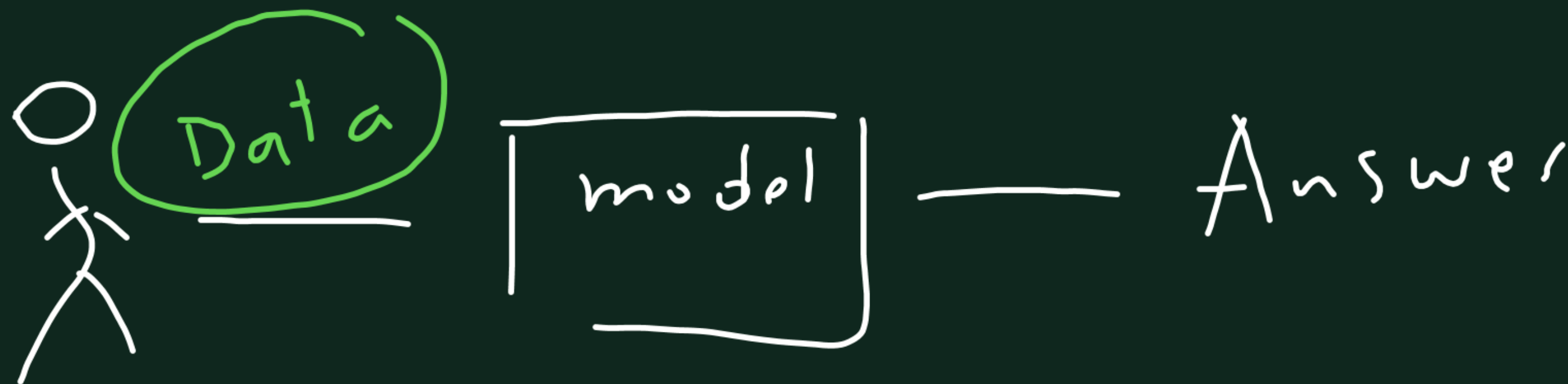
We start at 10:00

# Project Update

Due date: 22/09/24

~ 3 months

# Data (1)



## Data (2)

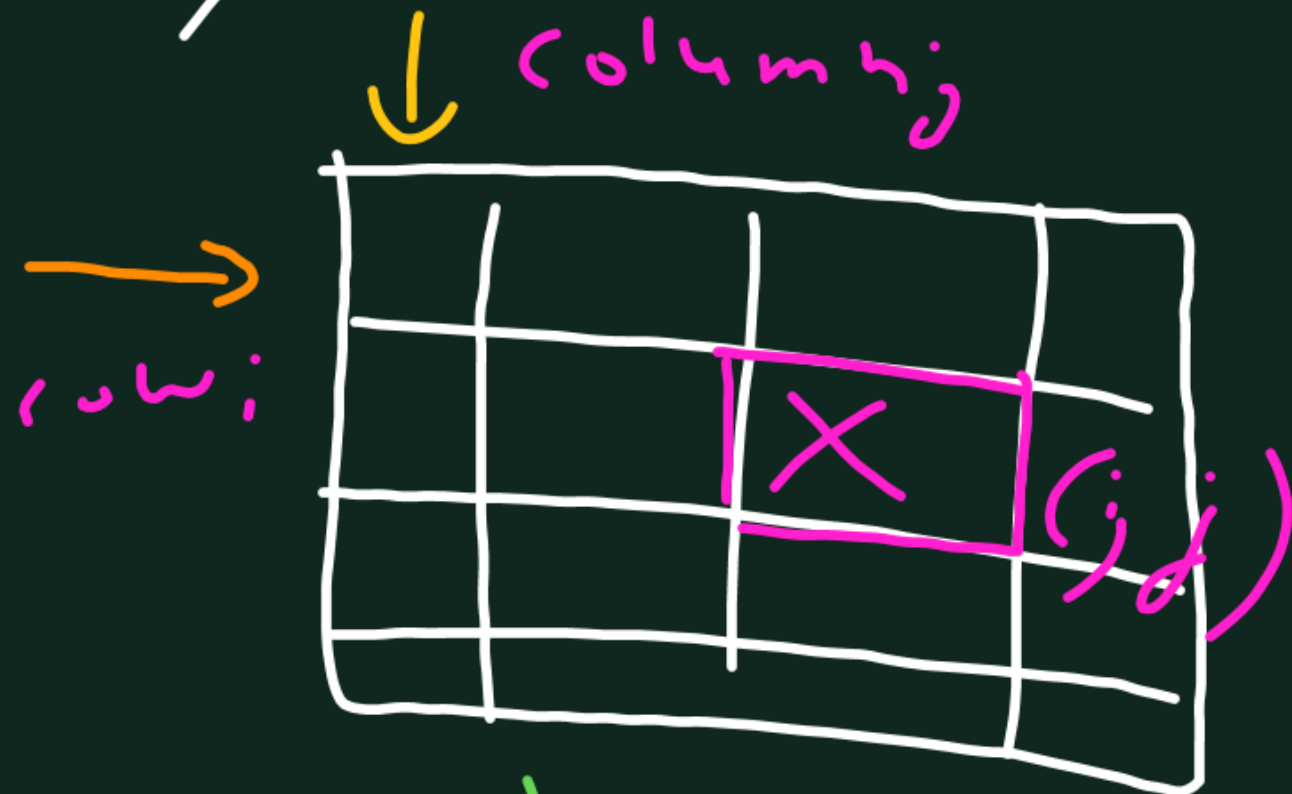
- Complex
- Disorganized
- Varied
- Multi-source

(DB / Excel /  
API / Web)

# Data (3)

To manipulate 'any' data :

- Heterogeneous
- Mutable
- Tabular



Dataframe

# Data (4)

Numpy

For mathematical  
operations

e.g.  $\text{pow}()$

Pandas

Manipulation  
of  
dataframes

Matplotlib

For  
plotting

SciKit-learn

For  
manipulating  
ML  
algorithms

# Data (5)

Pandas

Series:

Index

10

Mon

Tue

Wed

Thu

20

30

40

# Data (6)

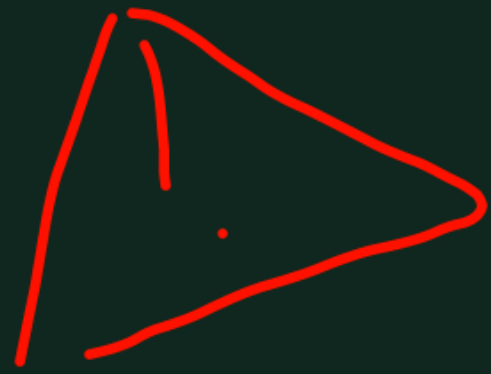
a	b	c
x	8	9
y	10	11

- cell
- Series (list)
- Dataframe (table)

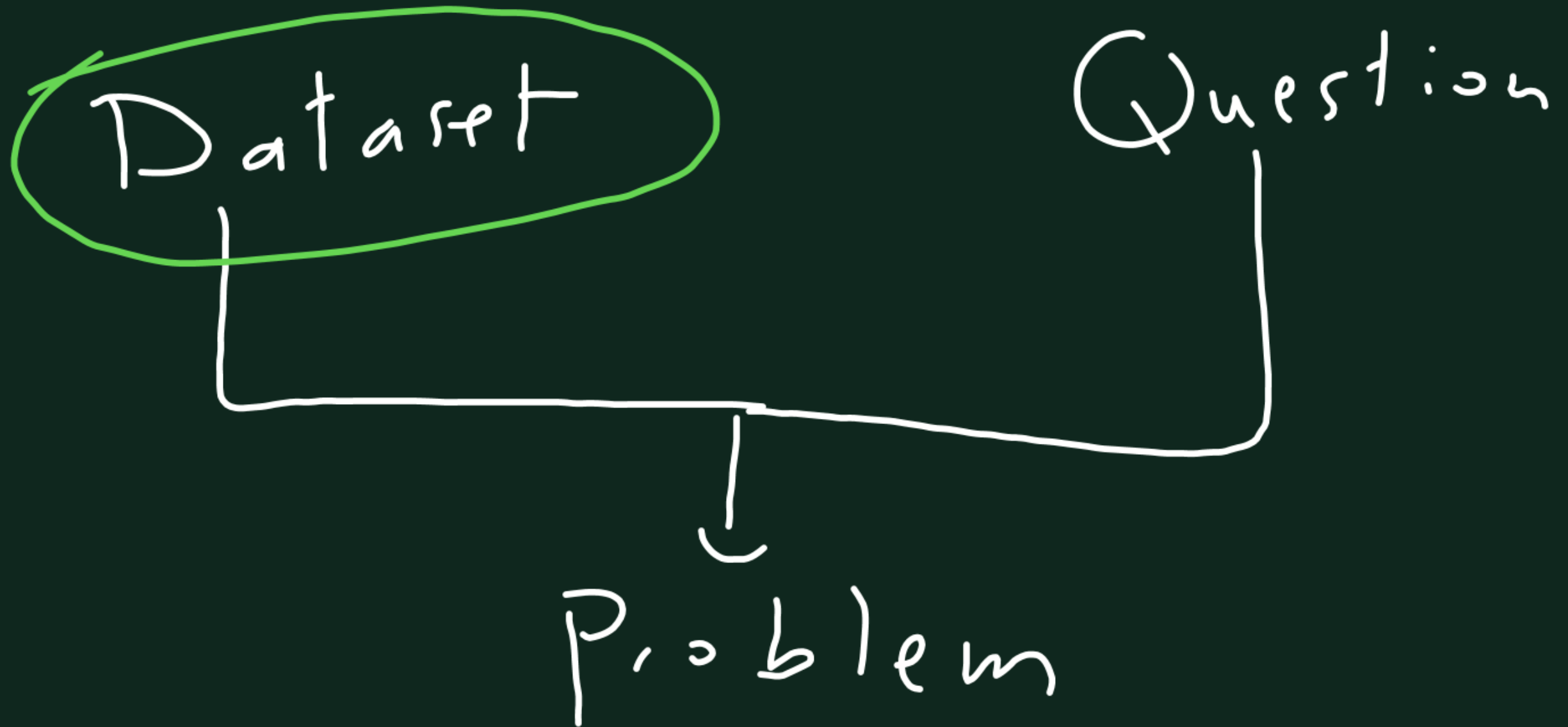


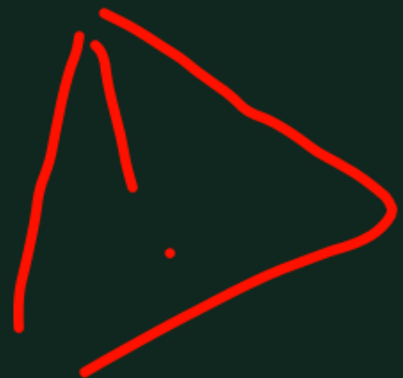
Break

Back at 11:30



# Machine learning (1)





# Machine Learning (2)

Problem:

Predict the value  
of the house?



Regression

Predict the ocean  
proximity of the house?



Classification

Lunch

Back at 14:00

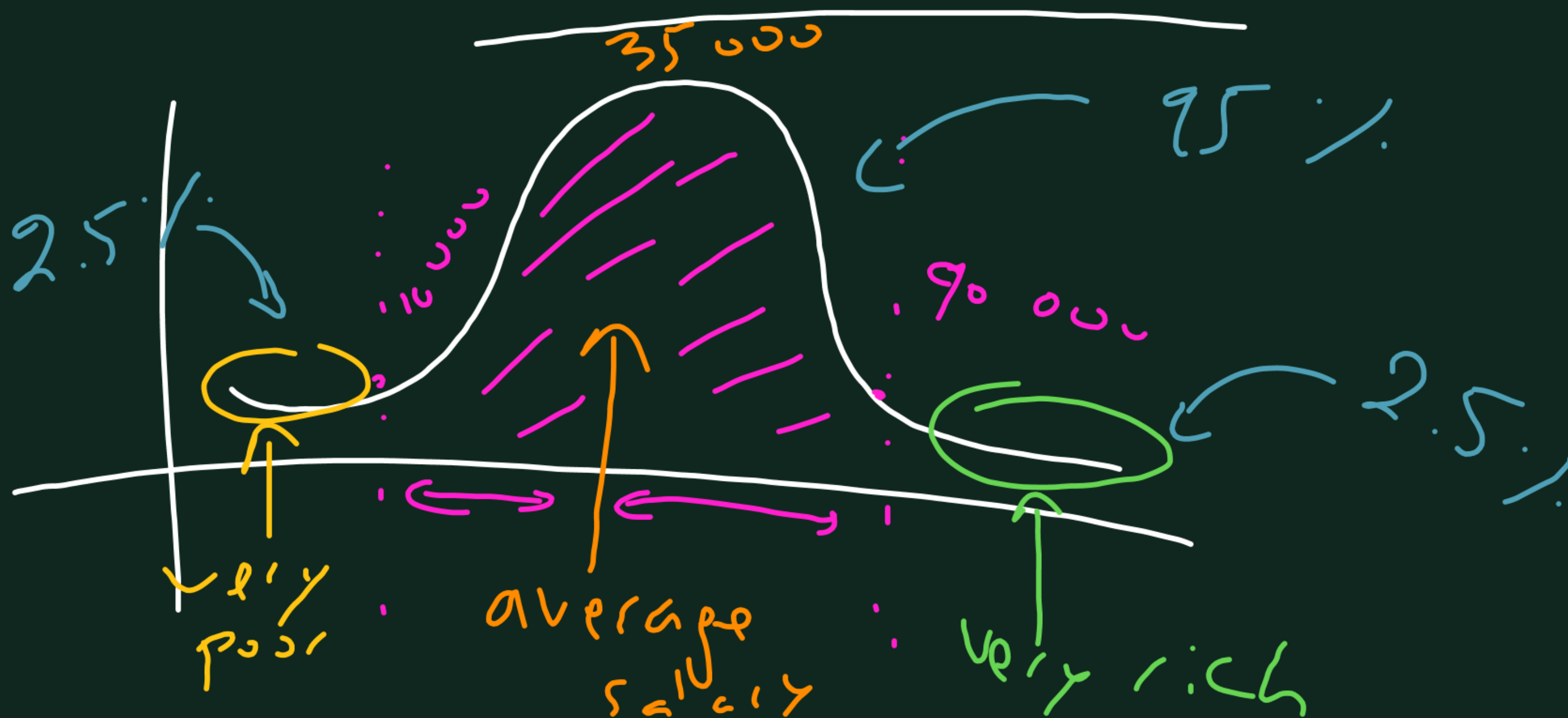
1%

## Outliers (1)

Age	Education	Gender	Salary
28 ✓	Master	M	60 000
56	PhD	(F)	128 000
27	Br.	M	35 000
28	No Diploma	M	300 000

Rich →

# Outliers (2)



Break

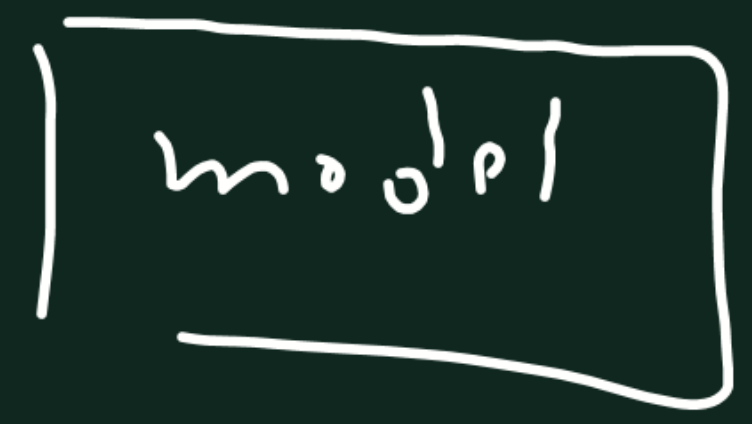
Back at 15:10

Education

Master's		
PhD		

NL

0	73	535
---	----	-----

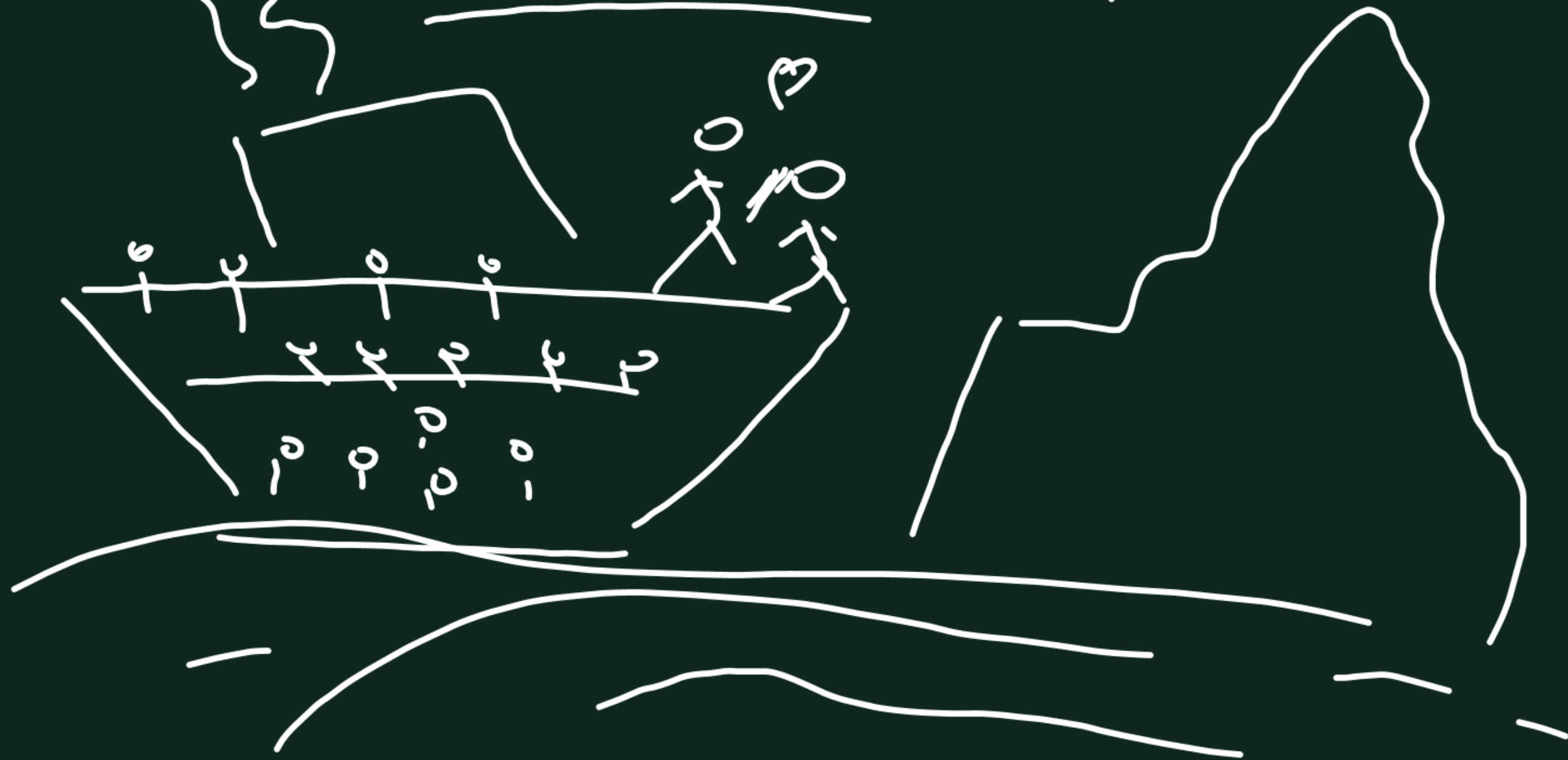


Output

- Human-readable dataset
- Machine-readable



# Titanic(1)



## Titanic (2)

"Can we predict a passenger's Survival?"

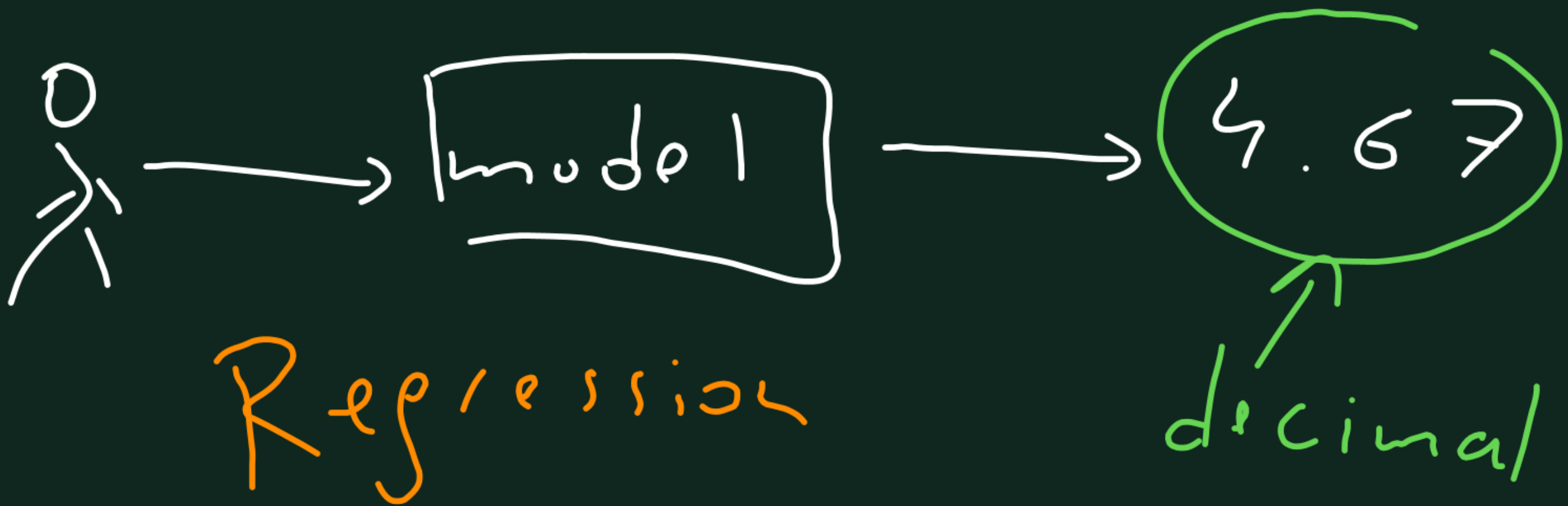


0  
(Not survived)

1  
(Survived)

# Book Project (1)

Can I predict a book's rating?



Categories

# Book Project (2)



# Book Project (3)



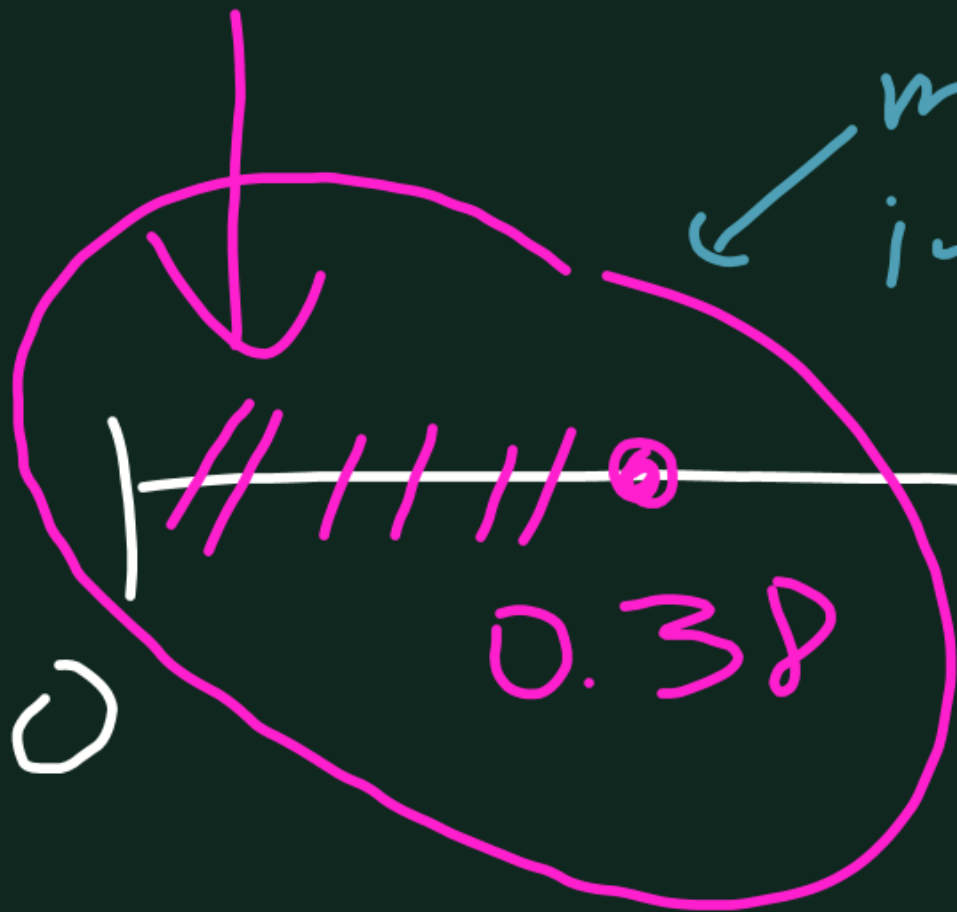
Titanic (3)

DEAD

Survival

Imbalanced  
Data

more  
info



less info